Individual Assignment 8

Please download the **Cereals.csv** dataset from Canvas.

This file contains data about a set of breakfast cereals.

Below are the column definitions for this sheet:

Cereal Product ID	Unique identifier for cereal
Cold	1 if cereal served cold, 0 if served hot
Calories	Calories/ serving
Protein	Grams protein/ serving
Fat	Grams fat/ serving
Sodium	Milligrams sodium/serving
Fiber	Grams fiber/serving
Sugar/ Carb Ratio	% of carbohydrates that are simple sugars
Potassium	Milligrams potassium/serving
Consumer Rating	Consumer rating, 0 being the worst, 100 being the best
American Cereal	1 if cereal sold in USA, 0 if not

To inform their marketing strategy, the firm producing these breakfast cereals wants to group them based on similarity. You will use this data to build a hierarchical clustering model in R that will produce groupings of similar cereals.

Open RStudio and create an R Markdown file. Save the .rmd file, ensuring that you save it in a file location that you can easily find again.

Each of the following items corresponds to a code chunk, or text to record in the R Markdown file. Please make sure to label each item using the numbers provided in this document.

Make sure to run each code chunk after you write it to ensure that there are no errors. Be careful about typos!

If you have not installed the caret, dplyr, and ggplot2 libraries yet, do so according to the instructions given in class before beginning this portion of the assignment.

Item 1, Loading Packages [Code Chunk]: Load the caret, dplyr, and ggplot2 libraries.

Item 2, Importing Data [Code Chunk]: Import the Cereals.csv file into an R data frame called "cereals".

Item 3, Pre-processing data [Code Chunk]: Normalize the data for hierarchical clustering, and save it in a new dataframe. Make sure that your normalized dataframe does not have the unique identifier column (Cereal Product ID) in it.

Item 4, Assessing Observation Similarity [Code Chunk]: Produce a distance matrix using the normalized data. Make sure the distance metric you are using is Euclidean distance.

Item 5, Generating Hierarchical Clustering [Code Chunk]: Using centroid linkage, run hierarchical clustering on the distance matrix. Plot the resulting dendrogram.

Individual Assignment 8

Item 6a, Setting Inter-Cluster Distance Cutoff [Code Chunk]: Plot the dendrogram again, but this time set a cut-off line on the graph for a desired inter-cluster distance of 0.8.

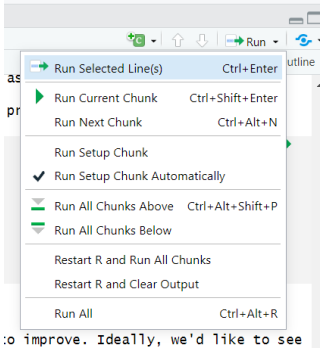
Item 6b, Setting Inter-Cluster Distance Cutoff [Text]: How many clusters are produced at this cut-off?

Item 7, Visualizing Clusters [Code Chunk]: Plot the dendrogram again. This time, visualize the number of clusters from indicated in 6b.

Item 8, Assigning Clusters [Code Chunk]: Using the number of clusters from 6b, assign each data record in the original data to a cluster. Save this in a new data frame called cereals_clustered.

Item 9, Evaluating Clusters [Code Chunk]: Run code to display how many records were placed into each cluster.

Click "Run All" to run all of your code chunks and check the output. Make sure that any text answers you have match this output. This is to ensure that your text answers match the output produced by R.



Save your file, then knit and export your R Markdown file as an HTML file. Upload the HTML file to Canvas to complete the assignment.

If you are unable to knit your .rmd file due to errors, make sure that you go back and test your code chunks individually. If you are ultimately unable to figure out how to solve these errors, save the .rmd file and upload that instead of the HTML file for partial credit.

If you are unable to find your exported HTML file, consult Canvas for instructions.